JLIS.it

DataCite and linked data

Jan Brase

The late Jim Gray from Microsoft Research has introduced the fourth science paradigms in his late work (Hey, Tansley, and Tolle). Thousand years ago science was empirical, describing natural phenomena. The last few hundred years saw a theoretical branch evolving, where models and generalizations were used to understand what was behind these natural phenomena, thus making the shift from the first paradigm to the second, with the scientists no longer being a passive observer, but actively trying to fins out, why things are like they are. The digital revolution in the last few decades allowed a computational branch to grow with the opportunity to use the developed theories to simulate complex phenomena. This was of course the shift from the second paradigm to the third, allowing the scientist to test in detail their theory against their empirical observation.

Today now with the next paradigm shift, we encounter what is labelled as enhanced science or eScience: Data intensive science that unifies theory, experiment, and simulation. This is what Jim Gray defined as the fourth paradigm. Now why is this important for libraries and what are the consequences of this for them? Libraries have a strong mandate of offering access to scientific information and knowledge. The German National Library of Science and Technology has a national mandate to provide scientific information to Academics and Industry in Germany. Secondly libraries have a long

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5493

tradition and experience in doing this, as they are doing so for thousand of years now. This makes libraries trustworthy organization that also have a tendency to be persistent. Especially in the digital age, where more on more information is only available in electronic formats this has become more and more important. While there is always a great risks that current projects and initiatives that create information will not longer be around after a decade, or to be more precise after the funding stops, the chance that the libraries will still be around are much higher. Following the paradigm shifts, information nowadays is more than article or books or any kind of textual information. If we take our mission seriously, we have to widen our mandate to any kind of information that might be relevant for our customers. This includes for example primary data, graphs videos, source code, power point slides, chemical structures among others. And other consequence directly effects us, the dramatic change in the definition of a library catalogue. Traditionally a library catalogue can be seen as a window to the library's holding, a structured summary of what can be brought easily to the shelf. Due to the growth of the internet in the last decades, this has slowly changed and more and more catalogues offer direct access to pdf-versions of document, but the principle has been the same throughout the centuries. Now in the fourth paradigm it becomes more and more impossible for a library to actively store all these kinds of information that are important for its user. Nevertheless the great chance with the growth of the internet is that the library does not have to store this information, when it is available somewhere else in the internet. The libraries job in the future is to know where the information is, if the content provider is trustworthy and to have a distinguished description of the content in its catalogue to offer the service of answering queries from user. In a nutshell, the library of the future should be able to answer the query of a user with the statement: «We do not have

what you are looking for, but we now where it is, and we can offer you a link to it». This implies many aspects: The library has to able to understand what the user is looking for. It has to able to have enough distinguished information about content in its catalogue to know what ideal results would be for the query. The library has furthermore to know where this content is stored and has to provide a persistent link to it to.
Today Technische Informationsbibliothek (TIB) is a global supplier for scientific and technical information, mostly traditional text-based documents. In the last years TIB as the German National Library of Science and Technology has started to actively open its catalogue or to be more precise its GetInfo portal to answer this challenge[1]. Nowadays you can use TIB's central information portal GetInfo as a search tool to access primary data, architectural models and chemical information.

Move beyond text - example

TIB is the German National Library for all areas of engineering as well as architecture, chemistry, information technology, mathematics and physics. It ranks as one of the world's largest specialist libraries, and one of the most efficient document suppliers in its subject areas. GetInfo, a portal for science and technology developed by TIB, bundles access to leading subject databases, publishing house offerings and library catalogues with integrated full text delivery. In doing this GetInfo offers a worldwide unrivalled supply of technical and natural scientific information. At present GetInfo is the only major library portal in Europe to include scientific datasets. The aim is to include all sorts of non-textual information into GetInfo.

The following two examples show data already included:

[1] https://getinfo.de

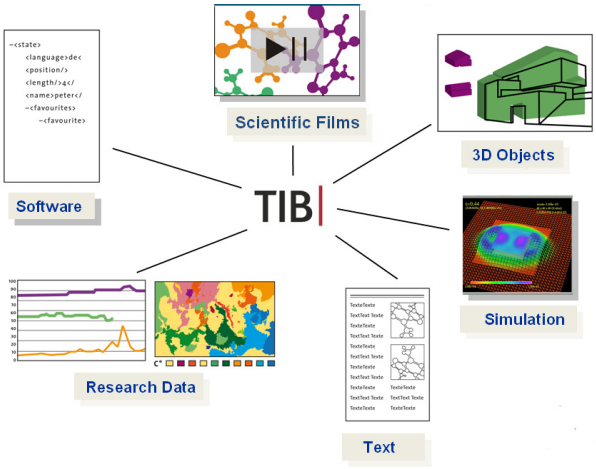

Figure 1: Contents and different ways to access to information

- Library catalogues are classical sources for information (Inger and Gardner). When querying for a certain topic, users might not be interested in only receiving all relebvant publications as a result, but also additional datasets collected by the corresponding researchers. The assignment of persistent identifiers allows this research data to become directly accessible through library catalogues. Nowadays a selection of more than 5.000 datasets that are part of scientific publications are directly accessible through GetInfo (Brase). When the persistent identifier of the dataset is resolved, the user does not directly download megabytes of data but is linked to a preview page where the data center provides metadata and download links to different parts of the data. This workflow is similar to the use of Digital Object Identifier (DOI) names in scholarly journals, where the resolution of a DOI name of an article directs you to a

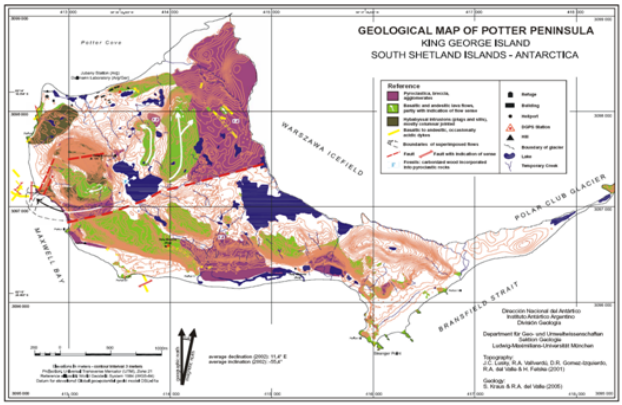

Figure 2: A geological map: non-textual information as rearch result.

publisher's page, including the metadata of the article.

- Another example of non-textual content in GetInfo can be seen in figure 2, where a geological map is displayed as a search result. Again the resolving of the identifier leads towards the preview provided by the specific data that includes the relevant information to this scientific object and displays the download link to the map.

As described earlier, the use of persistent identifiers for stabile linking between the catalogue and the external content is a fundamental requirement for the inclusion of non-textual information in our portal GetInfo. The registration of DOI names for scientific content especially scientific data has furthermore another consequence. Data sets than can persistently be linked to by DOI names become independently citable by other scientists.
Data integration with text is an important aspect of scientific collaboration. It allows verification of scientific results and joint research

activities on various aspects of the same problem. Only a very small proportion of the original data are published in conventional scientific journals. Existing policies on data archiving notwithstanding, in today's practice data are primarily stored in private files, not in secure institutional repositories, and effectively are lost This lack of access to scientific data is an obstacle to international research. It causes unnecessary duplication of research efforts, and the verification of results becomes difficult, if not impossible. Large amounts of research funds are spent every year to re-create already existing data. Encouragingly the "Brussels declaration on STM publishing"[2] focuses on dataset identification as a key element for allowing citation and long term integration of datasets into text as well as supporting a variety of data management activities. It would be an incentive to the author if a data publication had the rank of a citeable publication, adding to their reputation and ranking among their peers.

TIB developed and promotes the use of Digital Object Identifiers (DOI) for datasets. A DOI is used to cite and link to electronic resources (text as well as research data and other types of content). The DOI System differs from other reference systems commonly used on the Internet, such as the Uniform Resource Locator (URL), since it is permanently linked to the object itself, not just to the place in which the object is located. As a major advantage the use of the DOI system for registration permits the scientists and the publishers to use the same syntax and technical infrastructure for the referencing of datasets that are already established for the referencing of articles. The DOI system offers persistent links as stable references to scientific content and an easy way to connect the article with the underlying data. For example:

The dataset: G.Yancheva, N. R. Nowaczyk et al (2007) Rock magnetism and X-ray flourescence spectrometry analyses on sediment

[2] http://www.stm-assoc.org/brussels-declaration/.

cores of the Lake Huguang Maar, Southeast China, PANGAEA doi:10.1594/PANGAEA.587840 is a supplement to the article: G. Ycheva, N. R. Nowaczyk et al (2007) Influence of the intertropical convergence zone on the East Asian monsoon Nature 445, 74-77 doi:10.1038/nature05431.
Since 2005, TIB has been an official DOI Registration Agency with a focus on the registration of research data. The role of TIB is that of the actual DOI registration and the storage of the relevant metadata of the dataset. The research data themselves are not stored at TIB. The registration always takes place in cooperation with data centers or other trustworthy institutions that are responsible for quality assurance, storage and accessibility of the research data and the creation of metadata.

DataCite

Access to research data is nowadays defined as part of the national responsibilities and in recent years most national science organisations have addressed the need to increase the awareness of, and the accessibility to, research data. Nevertheless science itself is international; scientists are involved in global unions and projects, they share their scientific information with colleagues all over the world, they use national as well as foreign information providers.
When facing the challenge of increasing access to research data, a possible approach should be global cooperation for data access via national representatives:

- a global cooperation, because scientist work globally, scientific data are created and accessed globally;
- with national representatives, because most scientists are embedded in their national funding structures and research or-

ganisations.

The key point of this approach is the establishment of a Global DOI Registration agency for scientific content that will offer to all researchers dataset registration and cataloguing services. DataCite was officially launched on December 1st 2009 in London to offer worldwide DOI-registration of scientific data to actively offer scientists the possibility to publish their data as an independent citable object. Currently DataCite has 16 members from 11 countries:
The German National Library of Science and Technology (TIB), the German National Library of Medicine (ZB MED), the German National Library of Economics (ZBW) and the German GESIS – Leibniz Institute for the Social Sciences. Additional European members are: The Library of the ETH Zürich in Switzerland, the Library of TU Delft, from the Netherlands, the L'Institut de l'Information Scientifique et Technique (INIST) from France, The technical Information Center of Denmark, The British Library, the Swedish National Data Service (SND), the Conferenza dei Rettori delle Università Italiane (CRUI) from Italy. North America is represented through: the California Digital Library, the Office of Scientific and Technical Information (OSTI), the Purdue Univerversity and the Canada Institute for Scientific and Technical Information (CISTI). Furthermore the Australian National Data Service (ANDS) is a member.
DataCite offers through its members DOI registration for data centers, currently over 1.3 million objects have been registered with a DOI name.

References

Brase, Jan. "Using digital library techniques - Registration of scientific primary data". *Lecture notes in computer science* 3232. (2004): 488–494. (Cit. on p. 368).

Hey, Tony, Stuart Tansley, and Kristin Tolle, eds. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond: Microsoft Research, 2009. http://research.microsoft.com/en-us/collaboration/fourthparadigm. (Cit. on p. 365).

Inger, Simon and Tracy Gardner. "How readers navigate to scholarly content". (2008). http://www.sic.ox14.com/howreadersnavigatetoscholarlycontent.pdf. (Cit. on p. 368).

JAN BRASE, National Library of Science and Technology.
jan.brase@tib.uni-hannover.de

Brase, J. "DataCite and linked data". *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5493. DOI: 10.4403/jlis.it-5493. Web.

ABSTRACT: Science is global, it needs global standards, global workflows and is a cooperation of global players. But science is carried out locally by local scientists that are part of local infrastructures with local funders. DataCite is an international consortium, founded in 2009 of currently 17 institutions from 12 countries worldwide. Its mission is to allow a better re-use and citation of data sets. Over 1 million datasets have been registered with a DOI name as a persistent identifier, so they can be published as independent scientific objects to allow stabile citation of data. Citable data sets can be crosslinked from journal articles, their usage and citations can be measured therefore helping scientists gain credit for making their data available. DataCite offers a central metadata repository with additional linked data service for persistent access to RDF metadata.

KEYWORDS: DataCite; DOI; Data sets

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15